**Guidelines for Using the Public Data Release of the Community Investment Impact System
CDFI Program Data, 2004-2010**

These data user guidelines were prepared for the December 2011 CDFI Program Community Investment Impact System (CIIS) data release to explain the file structures, codebooks, data protocols and limitations and to facilitate the proper use and analysis of the data. Data analysts should note that this data release is far larger and more comprehensive than any previous public data release by the CDFI Fund.

I.   Introduction and File Structures

In order to facilitate the use of the several different types of CIIS data collected by the CDFI Fund, the CIIS data have been organized into four types of files:

  a. **Institution Level Report (ILR)** contains basic information on an organization's financials, staffing levels, and social impact metrics. The ILR is required of all CDFI Program awardees.

  b. **Transaction Level Report (TLR) File**: contains detailed financial information, including rates and terms, transaction status for all applicable CDFI transactions (e.g. delinquencies, charge-offs) and social impact metrics. Completion of the TLR financial and social impact report is required of many, but not all, CDFI Program awardees.

  c. **TLR Address File:** contains the physical location (city, state, zip, and Census tract) of all applicable transactions included in the TLR Financial file.

  d. **Portfolio Summary File** (**PSF):** contains summary data on loans originated and total portfolio outstanding for all CDFIs. This file synthesizes ILR and TLR data, allowing a comparison among CDFIs that are required to complete a TLR with those that are not required to do so.[1]

Because the TLR Financial and TLR Address files are both very large (over 400,000 records), they have been divided by year and loan/investment location:[2]

  a. **2004-2005** (one file, all states)
  b. **2006-2007 Part I** (Alaska-Kentucky)
  c. **2006-2007 Part II** (Louisiana-New York)
  d. **2006-2007 Part III** (Ohio-Wyoming and Ungeocoded)
  e. **2008 Part I** (Alaska-Louisiana)
  f. **2008 Part II** (Massachusetts-Wyoming and Ungeocoded)
  g. **2009 Part I** (Alaska-Louisiana)
  h. **2009 Part II** (Massachusetts-Wyoming and Ungeocoded)
  i. **2010** (one file, all states)

---

[1] CIIS reporting requirements for CDFI Program awardees have changed since CIIS was first launched in June, 2004. Originally, CDFIs were required to report their portfolio in an aggregate fashion as part of the ILR—the number and total dollar amount of loans/investments by type—but awardees are increasingly required to report portfolio information on a transaction basis in the TLR, including location, rates and terms, and community impacts. The Portfolio Summary File combines data from both types of files to enable a direct comparison between CDFIs. For example, consider two CDFI Program awardees that make microloans, of which only one is required to complete the TLR. The ILR would contain the total number and amount of microloans made by one CDFI, while the TLR would contain each individual microloan made by the other CDFI. The Portfolio Summary File would contain summary information on microlending by both CDFIs.

[2] Section V of this guidance discusses the limitations of address data in more detail. For purposes of dividing files, the first two digits of the **ProjectFIPSCode** were used to identify the state. Where no FIPS code was available, the transaction was labeled as "UNGEOCODED" and placed at the end of the alphabet (i.e. after Wyoming), regardless of where other variables situated the loan/investment.

II.   Limitations

Public data users must appreciate the following three major limitations:

a. **Data are only submitted to the Community Investment Impact System (CIIS) by CDFI Program awardees during the reporting period as required by their assistance agreements (typically three years).** The CDFI Fund estimates that CIIS captures data from approximately 15-20% of all certified CDFIs based on asset size. Certified CDFIs that are not awardees, or that have not recently received awards are not required to report any data to CIIS. Given the limited number of CDFIs reporting to CIIS it is not possible to draw conclusions about the CDFI industry as a whole based on these data. Similarly, it is not possible to draw definitive conclusions about overall CDFI activity in a specific location (such as an analysis of CDFI coverage) because CIIS does not capture lending and investing activities by CDFIs that are not recent awardees.

b. **The number of CDFIs reporting to CIIS differs each year as awardees enter and exit the reporting system under the terms of their assistance agreements.** Therefore year-to-year comparisons may be misleading unless a single cohort is also followed over time. For example, based on the CDFI Fund's analysis of the 2003-2005 there were 86 CDFIs that reported in all three years of this time period.

c. **Data points in the TLR Financial file represent all transactions reported in a CDFI's portfolio, not just those transactions originated in a given year.** Users should note that the same loan may be reported in the TLR years after it was originated; thus, if the principal balance or loan status has changed, the resulting CDFI activity on a year-to-year basis will change and such changes must be taken into account to avoid misstating the level of activity. To compare TLR data from one year to another, transactions should be grouped by the time period in which they were originated, rather than the year in which they are reported in CIIS.

For example, a business loan originated in 2007 may have created 50 jobs, according to a CDFI's 2007 TLR. In the 2008 TLR, the loan is reported again to note the new principal balance outstanding, and in the 2009 TLR, the loan is reported one final time to note that the balance has been paid off. The 50 jobs created by this loan would be reported for all three years, so that analyst should count the outcome only once to avoid inflated estimates (e.g. an incorrect count of 150 jobs created rather than a correct count of 50).

Identifying the loan by its origination date (**DateClosed** in the TLR Financial file), and not by the reporting year, allows the correct number of jobs to be credited to this transaction. The origination date and reporting year should be aligned based on a notional fiscal year. (If the Federal fiscal year of October 1-September 30 were used as the baseline, a loan originated in September, 2007 would be reported in the 2007 fiscal year and a loan originated in October, 2007 would be reported in the 2008 fiscal year. Any records linking these origination dates to another fiscal year should be deleted for data analysis purposes.)

III.   Data Suppression and Privacy

To preserve privacy and confidentiality, certain data elements are omitted or suppressed, while other data fields may be modified. This ensures compliance with federal privacy laws and regulations. Further details are available in the CDFI Fund's guidance on releasing data to researchers, published on the CDFI Fund website:

Data Release Protocol, Institution Level Report
Data Release Protocol, Transaction Level Report

IV.   Treatment of Missing Data and Added/Dropped Questions

The CDFI Fund has prepared codebooks to assist users in the analysis of these data. The four CIIS data files contain four types of variables: key (identifying), categorical, numeric, and combined. Key variables, such as fiscal year or organization ID, have no missing values. All other types of variables may include missing values, and the reporting of missing data differs by type.

Categorical Variables
Categorical variables list the total number of observations for each year, the various categories that are potential responses, and the number of missing values. The number of actual, counted responses is the N minus the number of missing values.

		2004	2005	2006
Guarantee	**N**	6,486	32,815	55,572
Type of Loan Guarantee	**NONE**	429	1,551	8,281
	OTHER	16	785	1,580
	OTHERGOV	1	2	12
	PERSONAL		57	6,597
	SBA			58
	USDA			23
	(MISSING)	6,040	30,420	39,021

Questions and responses vary slightly from year to year. In the above example, a zero response for "PERSONAL" in 2004 does not mean that no borrowers had personal loan guarantees. The response is simply unavailable. Additionally, a question that was added to the TLR after 2004 will only have "missing" values for the prior years. For example, in the TLR Financial file, the acceptable responses for Loan Purpose changed during the 2005 reporting year: "BUSWORKCAP" (business loan-working capital) and "BUSFIXED" (business loan-fixed) were replaced with two new categories, "BUSINESS" and "MICRO." The changes in categories reflect changes in measurement, not in actual CDFI activity. Microloans were made prior to the change, and working capital loans are still being made; only the reporting options have changed.

Numeric Variables
For numeric variables, the N is not the total number of observations, but the total number of non-missing responses. For example, in 2004, the mean total project cost was $833,715 across 810 observations. Zero values were not treated as missing.

		2004	2005	2006
TotalProjectCost	**N**	810	5,748	5,136
Total Project Cost	**MIN**	0	477	500
	MAX	38,200,238	187,063,471	187,063,471
	MEAN	833,715	756,474	1,152,013

Combined Variables
Combined variables contain both categorical elements (e.g., DK=Don't Know and NA=Not Applicable) and numeric elements. For combined variables, the N listed is not the total number of observations, but the total number of numeric responses. All other responses (DK, NA, or missing) are not counted as part of the N.

		2004	2005	2006
HousingUnitsSale	**N**	72	679	1,531
Housing Units - Sale	**DK**		81	670
	NA	5,769	27,663	16,129
	(MISSING)	645	4,392	37,242
Housing Units - Sale	**MIN**	0	0	0
	MAX	928	210	700
	MEAN	16	5	5

V. Tips for Effectively Using Investee/Project Address Information

Project address information is included in the Address Transaction Level Report (TLR) file. This information is useful to analyze the distribution of CDFI loans and investments, after taking into account the data limitations noted above.

With few exceptions, all address information has been left as is, including zip codes and FIPS codes that may or may not be valid or accurate. To adequately account for transactions in a given unit of geography (city, state, etc.), the user should verify that address information is consistent across variables before making conclusions. When tabulating data at the state level, it may be useful to summarize data by state, and also by the first two digits of the FIPS code (which identifies the state). Comparing these two sets of results may reveal slight differences. Or, if looking at transactions in a particular city, it may be helpful to also search for the zip codes or Census tracts that make up that city. TLR Address data should be treated with caution. Typographical errors and other inaccuracies do occur, and given the sheer number of transactions (over 400,000), it is difficult to identify or correct all of them.

A few examples demonstrate the challenge of address consistency in these files:

1) When searching by city for all transactions in Little Rock, the list includes the following:

projectcity	projectstate	Project ZipCode5	Project FipsCode
LITTLE ROCK	AR	72206	05119000700
LITTLE ROCK	SC	29567	45033970100

Both of these addresses are consistent. The zip codes match up with each location (72206 for Little Rock, AR and 29567 for Little Rock, SC). Additionally, the first two digits of the FIPS code correctly identify each state (05 for Arkansas, 45 for South Carolina). An accurate search would need to incorporate not only the city, but the state as well. Searching only for "Little Rock" would include South Carolina transactions as well as those in Arkansas.

2) However, searching for all Chicago transactions in the state of Illinois would not identify the following transaction:

| projectcity	projectstate	Project ZipCode5	Project FipsCode
CHICAGO	TX	60618	17031160900

The zip code appears to be for Chicago, and the first two digits of the FIPS code (17) represent Illinois, but the state is incorrectly labeled as Texas. Searching by city and state separately, or searching by city and FIPS code together, would safeguard against missing such cases in a search.

VI. Relationships Among Files

Some analyses may require linking together files or, at the very least, organizations. For example, suppose you are interested in the total assets of each CDFI that has reported loans in Colorado. Locations of loans/investments are only in the Address TLR (Transaction Level Report) files, whereas assets are only in the ILR (Institution Level Report) file. As shown in the graphic below, records in the two files are linked by the fiscal year and the organization ID of each CDFI. Once the transactions in Colorado have been identified, the fiscal years and organization IDs can be linked back to the ILR to determine the asset size of each institution.

The Institution Level Report (ILR) can be linked to any of the other files by using the combination of fiscal year and organization ID. The Financial Transaction Level Report (Financial TLR) can also be linked to the Portfolio Summary File in the same manner. However, the two types of Transaction Level Report files (Financial and Address) are linked in their own way by a variable named **projectID**. This variable alone enables the two files to be linked without requiring any other fields.

RELATIONSHIPS AMONG FILES
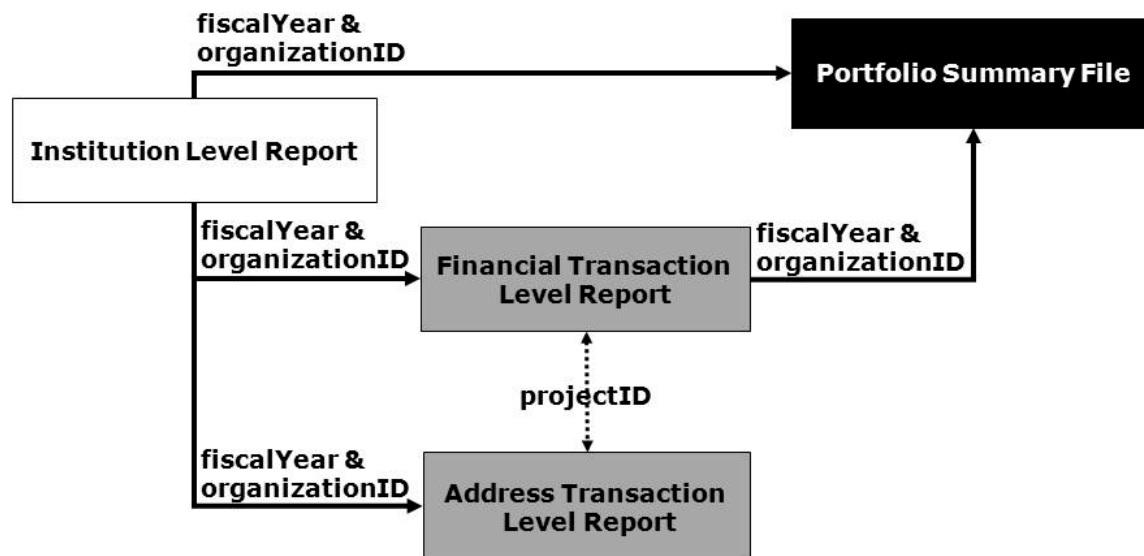
Note: Because a single transaction may occur at multiple addresses (e.g. a residential real estate project at scattered sites in the same city), merging these two files will occasionally result in the same transaction being counted twice. The effect may not be very large, but the number of transactions reported would change.